APPENDIX A

Statistical Informatics

Procedures for Analyses of Array Data

Introduction

Array-based expression analysis (ABEx) is projected to find increasing application in drug discovery and diagnostics. Although ABEx tools remain complex, our belief is that we are entering a period in which a rapid evolution of skills and commercial tools will favor increased application of this technology.

Some of the ABEx evolution will lie in materials, processes, and instrumentation. Advances will be made in microfabrication, hybridization procedures, arraying and detection. Another evolving aspect relates to the ways in which meaning is extracted from the arrays - informatics. Questions regarding meaning are posed at two levels.

- **Validity of observations.** How do we verify that our observations are real?
- **Consequential validity.** Do our observations have meaning in terms of biological consequences for the organism? For example, can we identify consistent "expression patterns" in families of genes that show similar activity under conditions of interest? Can alterations in expression of specific genes be linked to translational/post-translational events?

All ABEx users will face both these questions. Consequential validity is the goal of the research. Verified observational validity is a precondition to drawing conclusions regarding consequential validity.

We believe that complete ABEx systems must both detect and specify validity for high density arrays of hybridization data. Therefore, we are developing tools for "statistical informatics". Statistical informatics (SI) is a set of analytical procedures that provide reliability estimates of ABEx data points. Statistically reliable data are more likely to be valid.

Classes of Expression Study

ABEx studies fall into three general classes.

Single condition:	Expression in a single condition, without comparison to control.
Diagnostic:	Expression in a single condition, with comparison to a standard control.
Comparative:	Direct comparisons of expression across conditions.

Single condition studies (e.g. Pietu et al., 1996) make reports of the order "We found X highly expressed sequences in this tissue, some of which are not expressed in other tissues". These types of reports are becoming less common, because it is very difficult to establish a causal link between an expression observation and a tissue condition.

Diagnostic and comparative studies perform direct comparison of specimen conditions. In the diagnostic case, the comparison is with an independently standardized control condition. In the comparative case, multiple conditions are included within an experiment.

Expression Data

All classes of ABEx study must yield reliable expression values. Note that the term reliable, as used here, is analytical as opposed to biological. Reliable expression values are those which have a specified (and preferably low) error variance. There are various strategies used to decrease error variance in array data.

- **Multiple spotting.** Replication (using multiple instances of each probe) paradigms have long been used to minimize effects of variation within an assay. By using multiple replicates of an assay and

taking an average, or excluding highly variable cases, we achieve a more reliable result than if we take a single case.

- **Fluorescent label.** A very large improvement in data quality can be obtained by moving from isotopically labeled specimens on nylon membranes to fluorescently labeled specimens on glass substrates.
- **Comparing conditions using multiple labels on the same probe.** Multiple fluorescent labels minimize error variance by allowing direct comparisons between different conditions (e.g. cancer vs. normal) hybridized to a single probe (e.g. DeRisi et al., 1996; Shalon, Smith and Brown, 1996). In this case, data are expressed as ratios between conditions.
- **Reference the mean, median, or a set of reference genes.** Each member of the array may be referenced to some global parameter. Theoretically, this process will minimize intersample variation by removing reliance on absolute intensity values.
- **Match-mismatch pairs.** Each sequence in the array has a companion that differs, usually by one base pair. Data are expressed as a ratio of the "perfect match" to the mismatched sequence or as a subtracted value (match - mismatch). Subtraction removes nonspecific hybridization and background (which should be the same under both conditions), and normalization provides an internal reference for the probe in question.

Once we have produced a body of data, the next step is specification of how much alteration in expression is meaningful. For example, one sees statements such as "2:1 alterations in expression are detectable". The precise justification for this type of statement varies. The most common approach uses an estimate of variability derived from reference genes. The library contains a set of reference or "housekeeping" genes, which are known to hybridize. Variance in this set is used to establish a variance criterion for other members of the array.

A New Procedure

We propose a statistical procedure, which we call "statistical informatics" (SI), for analyses of ABEx data. SI includes two major components.

a) Deconvolution of distributions. If the array data include contributions from two or more distributions (e.g. signal/nonsignal, multiple fluors), we deconvolve those distributions into distinct probability density functions. This allows discriminating of hybridization signal from nonsignal, and/or discriminating contributions of one label from another;

b) Reliability of expression values. Some of our observed expression values are good estimates (reliable). Others are heavily influenced by error (unreliable). For any expression value, we calculate reliability.

Advantages of SI include:

- accept data generated using any variance reduction strategies;
- model-based, as opposed to using reference materials created with the array;
- are simple to use, in that generic arrays can be analyzed;
- provide an objective method for calculating the reliability of each data point.

Discriminating Distributions

Many ABEx data arrays are composed of multiple distributions. For example, a hybridization data set provides both signal and nonsignal elements (Figures 1,2). Discrimination of nonsignal is necessary so that we can make meaningful comparisons of expression (signal:signal), while avoiding spurious comparisons (any that include nonsignal).

Figure 1: Frequency distribution of a simulated array, showing a mixture of both signal and nonsignal assays. Background has a mean of zero, and varies about that value. Therefore, there are both positive and negative values in the distribution. This type of distribution is typical of arrays on nylon membranes.

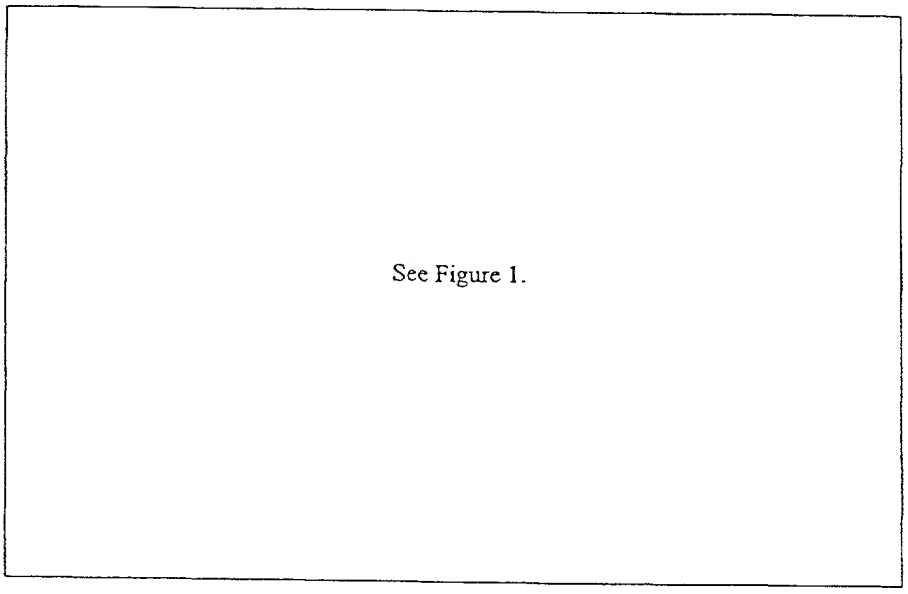
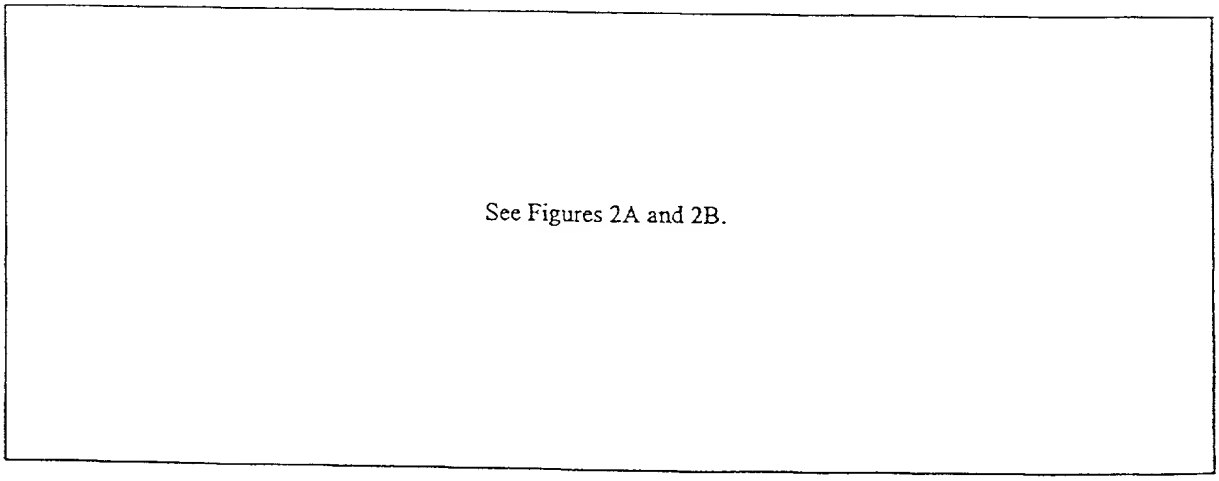


Figure 2. Distributions of signal and nonsignal generated from the data set in Fig. 1.



The Modeling Process

Step 1: Describe probability density functions for the two distributions, using modeling. We create a set of descriptors, that specify the nature of each distribution. To create these descriptors, we make an assumption that each distribution originates from a specific probability density function (pdf) which can be estimated from four parameters - mean, variance, proportion of the mixture, and class (e.g., Gaussian, gamma). A well-accepted method for deriving mean, variance, and proportion of mixture from mixed distributions is maximum likelihood estimation (MLE). Other methods could be used.

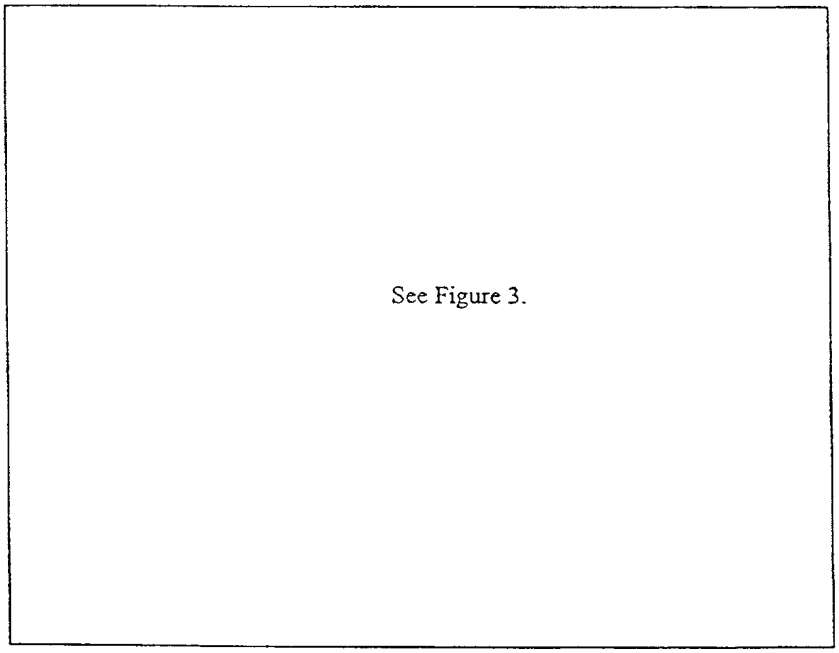
Definitions

Maximum likelihood method: We ask, "How likely is it that we would have obtained the actual data given values (generated by software or by the user) for four parameters for each distribution (mean, variance, proportion of mixture, and distribution class?" (e.g., Gaussian, gamma). The MLE procedure estimates the likelihood of obtaining the actual data given the initial values, and then proceeds to evaluate this likelihood given slightly different values. Iteration continues until it arrives at a likelihood that is at its maximum or until a predefined iteration limit is reached.

Probability density function: A curve (e.g., Gaussian) defined by a mathematical equation. Probabilities for ranges of values (e.g., $x > 100$; $x < 500$) can be derived based on area under the curve.

The MLE procedure generates pdfs for the signal and nonsignal distributions (Figure 3). These distributions include areas that are unambiguously part of one distribution or another. They also contain an area of overlap, and it is in this overlap area that our process operates to assign the origin of data points.

Figure 3: Probability density functions of the signal and nonsignal distributions, showing the region of overlap. Within this region, our process assigns hybridization values to distribution of origin.



Step 2: Use the probability density function to assign hybridization values to their distribution of origin. For any hybridization value, we can determine the probability of obtaining a value that large or larger from the nonsignal distribution or that small or smaller from the signal distribution. In this way, we obtain two probabilities (one that the value came from the nonsignal distribution and one that the value

came from the signal distribution). Comparing the two probabilities tells us which distribution is the more likely originator of the data value.

Consider the values reported in Table 1, which were taken from the simulated data discussed in Appendix A. There are three things to note:

1. Higher values are less likely to have come from the nonsignal distribution (see Column 2) and more likely to have come from the signal distribution (see Column 3).
2. The probabilities in Columns 2 and 3 show which of the two distributions is more likely to be the origin of a particular hybridization value. For example, the probability that a value of 40 or greater came from the nonsignal distribution is .2107. The probability that a value of 40 or lesser came from the signal distribution is .0995. Our procedure establishes that a value of 40 is more likely to have come from the nonsignal distribution.
3. A criterion value for signal and nonsignal hybridization can be obtained from the probability function. In our example, a value less than or equal to 49 is categorized as nonsignal and greater than 49 is categorized as signal.

Table 1. Probabilities of origin for various hybridization values.

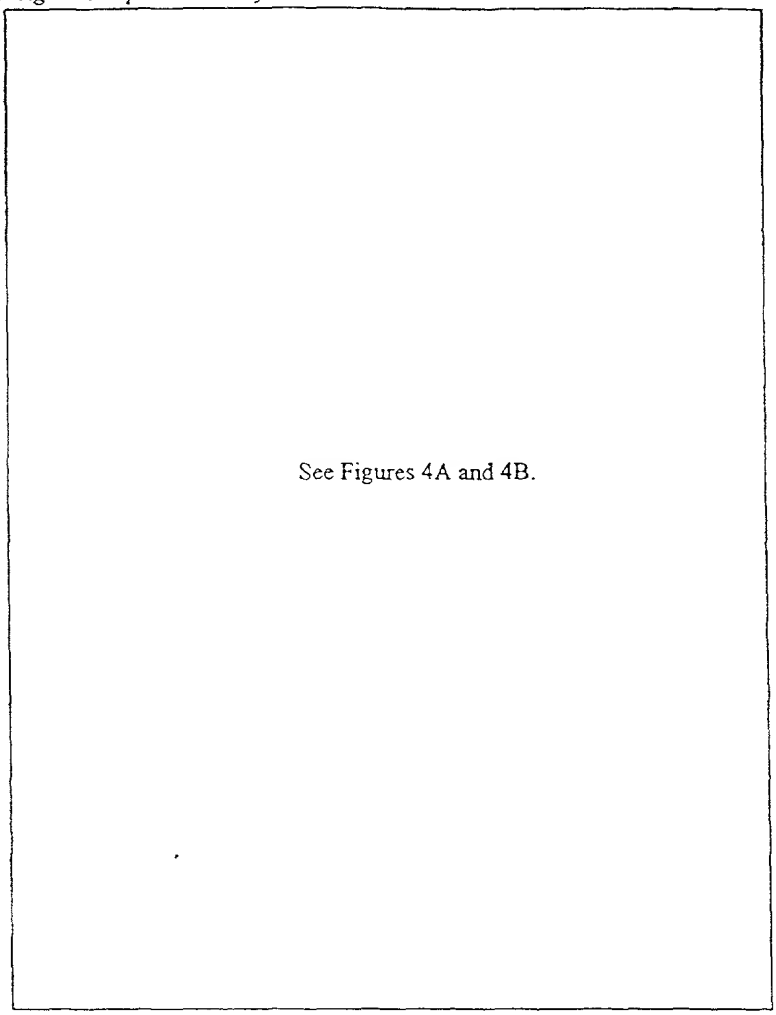
Value	Probability of Originating from the Nonsignal Distribution	Probability of Originating from the Signal Distribution	More Likely Originating Distribution
40	.2107	.0995	Nonsignal
45	.1740	.1258	Nonsignal
49	.1493	.1482	Nonsignal
50	.1436	.1540	Signal
60	.0980	.2148	Signal
70	.0669	.2788	Signal
78	.0493	.3308	Signal

Step 3: Test Goodness of Fit. The present invention creates models which purport to describe real data. We can evaluate the models using a goodness of fit parameter based on the chi-square statistic. The test can be automated, and the software flags cases in which the modeling results in a bad fit.

When Modeling is Appropriate

The modeling procedure assumes that the array of hybridization data points can be parsed into multiple distributions, each with sufficient members to allow accurate modeling. This is usually the case with nylon arrays, which contain large nonsignal components (Figure 4). Many glass arrays are quite different in nature. The background tends to be much lower, and the signal to noise higher. Therefore, it may not be possible or necessary to model a nonsignal distribution for very clean arrays. In the case of a clean glass array with a single label, we can assume a single (signal) distribution, dispense with the modeling, and use a simple signal criterion to discriminate usable assays (e.g. assays with a signal to noise ratio >3:1).

Figure 4: Distributions of data showing two nonsignal proportions. Top is a 32P-labeled Clonetech Atlas array on nylon. There is a large nonsignal component. Bottom is a Cy3-labeled glass microarray (muscle tissue) The nonsignal component is very small.



Modeling a Real Specimen

To summarize the situation to this point:

We have demonstrated that modeling works well with a theoretical distribution.
We have shown that membrane arrays have the properties of the theoretical distribution.
Clean glass microarrays may not have enough nonsignal points to allow modeling.

Will the modeling be useful with glass? To answer this question, we examined some microarrays that are less clean than our excellent lymphocyte library array. In fact, these arrays have many of the properties of membranes (Figs. 5,6). Therefore, the modeling will be useful with a broad variety of arrays, including fluorescent microarrays.

Figure 5: A Cy3 fluorescent microarray image integrating three replicates of a spinal cord library. The dim red dots represent nonsignal. The brighter red dots fall into the area of overlap where the modeling might assign them to either signal or nonsignal. Other colors are unambiguously signal.

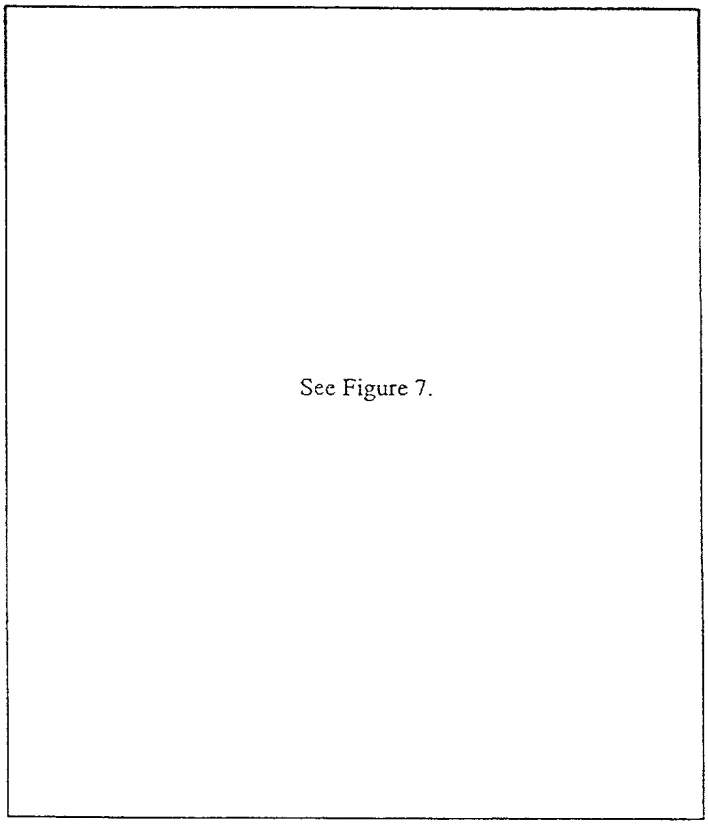
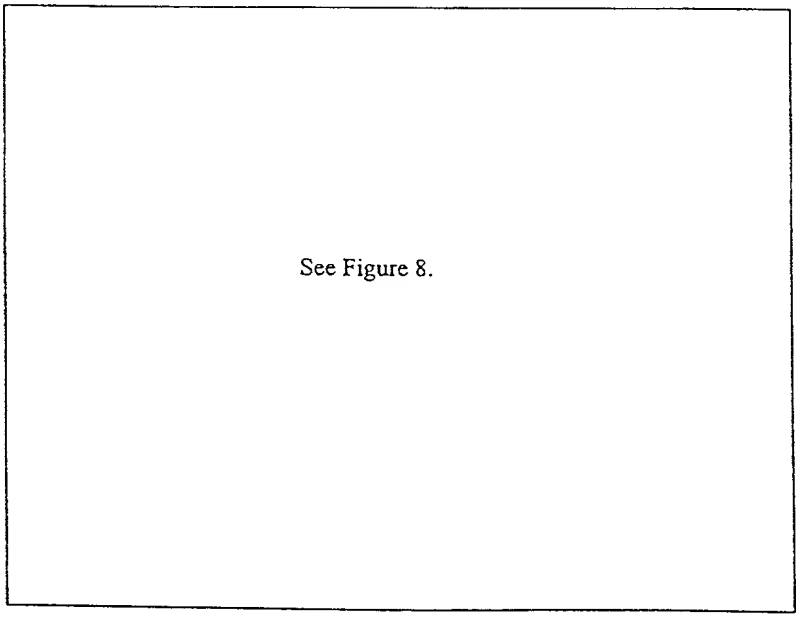


Figure 6 Modeling of the array in figure 5. The red lines show the distributions of signal and nonsignal. The blue shows intensity bins. The green line represents the modeled fit to the actual data The model does not differ, significantly, from the data ($\chi 2$ test).



Summary of Distribution Modeling

We use modeling procedures to deconvolve a data matrix into two or more probability density functions. Hybridization data are then assigned to the most likely distribution of origin. Advantages of the modeling are:

- no need to create reference arrays to estimate nonsignal;
- objective assignment of hybridization values to signal or nonsignal distributions, to one label or another, or to any other deconvolved distributions.

The process can include a goodness of fit test, which alerts us if the outcome of the modeling is suspect.

Reliability and Confidence Intervals

Any hybridization assay is an estimate. That is, if we repeat the assay a number of times, we will obtain values which vary about a mean. All of these observed values estimate a true hybridization value. Some assay values are reliable estimates of the true value, and others are not. It is useful to specify the extent to which any given expression value is reliable.

Confidence intervals bracket a true value. In defining confidence limits for hybridization, we use the observed values as estimates, and generate ranges around the estimates. Given an observed value of X, and an estimate of the reliability of the observed value, we can give a range within which the true hybridization value estimated by X should lie. This range is stated with a particular confidence (e.g., > 95%).

We can also use the range data to specify our confidence in differences between assay values or expression ratios. If the ranges overlap, we have low confidence in the differences. If the ranges do not overlap, we have high confidence.

The Process: Replicates Are Present

If replicates are present, measurement error can be determined, directly. The additional advantage of replicates is that error associated with an average is decreased by a factor of $1/\sqrt{n}$ where n is the number of replicates.

Step 1: Identify highly unreliable assays using estimates of variance derived from the replicates. Estimates of variability across replicates will vary from assay to assay. If they vary too much, the assay should be discarded. How do we set the criterion for discarding an assay?

We examine the variability of the variability. From this, we can identify replicates whose variability exceeds a value. The value is determined by calculating the variance of the variance values, and setting an objective variance criterion (e.g. 3 SD units) to indicate outliers.

Step 2: Determine error estimates for the acceptable assays using either standard error of the mean or coefficient of variation. True assay values are estimated by the mean of the replicates. The process can use either the standard error of the mean ($\hat{\sigma}_{\bar{x}}$, eq 1) or the coefficient of variation for the mean ($CV_{\bar{x}}$, eq. 2) to estimate assay error from the replicates.

Equation 1. Standard error of the mean of the replicates for a given assay.

$$\hat{\sigma}_{\bar{x}} = \hat{\sigma}_x / \sqrt{N},$$
where $\hat{\sigma}_x$ = the standard deviation of the replicates,
and N = the number of replicates.

Equation 2. Coefficient of variation for the mean of the replicates for a given assay.

$$\text{Percentage CV}_{\bar{x}} = 100(\hat{\sigma}_{\bar{x}}/\bar{x}).$$

In the case of additive error (e.g., 100 ± 10, 1000 ± 10), the standard deviation is the best estimator of variance around each data point. The absolute value of error remains constant.

In the case of proportional error (e.g., 100 ± 10, 1000 ± 100), the coefficient of variation is a more useful measure of variability. The standard deviation changes proportionally to the magnitude of the measurement value.

Raw score hybridization assays will, typically, present proportional error, whereas log transformed assays will present additive error. The appropriate statistic is chosen on that basis.

To summarize the process, we obtain an average SD or CV for the replicates in the entire array. We then use that average in the next step.

Step 3: **Calculation of confidence intervals.** Error estimates for the assays allow us to construct confidence intervals around each assay. The higher the confidence we wish to have, the broader the range that brackets the true value. The range of possible values at a particular *confidence level* is called a "*confidence interval.*" Ninety-five percent and 99% confidence are typical confidence levels.

Confidence level: The probability that our range includes the true value.
Confidence interval: The actual values of the range.

Step 4. **Using confidence intervals for comparisons among assays.** The measured values for any two assays will almost certainly differ from each other. These differences may simply reflect the effects of measurement error or they may reflect actual differences between the true values. We use confidence intervals to give probabilities that an observed difference is real.

If the confidence intervals of two assays do not overlap, we have confidence at the chosen level (e.g., 95 or 99%), that the true values of the assays differ from each other. If the confidence intervals overlap, we do not have confidence that the true values differ.

The advantages of our procedures are:

- Error is calculated from replicates, using standard statistical procedures.
- The confidence intervals are calculated directly from the array data.
- Confidence intervals are stated, using objective criteria.
- Expression comparisons are given with a probability of error.

The Process: Replicates Are Not Present

If replicates are not present, statistically derived estimates of reliability are unavailable. However, we still need error estimates to construct confidence limits. These error estimates are created in various ways. We can build some replicated assays into the array, and estimate error from these (e.g. DeRisi et al., 1996). Alternatively, the user can enter some error value that is characteristic of his data sets.

Once an error estimate has been specified, confidence limits can be calculated and comparisons among expression values can be specified with probabilities.

Analyzing Differential Expression

Most modeling processes require large numbers of data points. Usually, comparing hybridization values across conditions does not provide large numbers of differentially expressed assays. Rather, there tends to be a large number of assays with similar ratios (usually 1:1), and a relatively few cases of differential expression (e.g. 4:1). This creates difficulties for accurate modeling.

Fortunately, we can take advantage of some properties of the ratio to conduct distributional modeling that do not require large numbers of data points.

The Process: Measurement Error Known

Generate confidence intervals for expression ratios using replicates or user entry to estimate measurement error. If we have estimates of the measurement errors associated with the numerator and denominator of a ratio, it is a simple matter to estimate the measurement error associated with the ratio.

Equation 3 Percentage error for hybridization ratios (replicates present).

$$\text{Percentage error A/B} = 100\sqrt{\left(\frac{\hat{\sigma}_{\bar{x}_A}}{\bar{x}_A}\right)^2 + \left(\frac{\hat{\sigma}_{\bar{x}_B}}{\bar{x}_B}\right)^2}$$

where $(\hat{\sigma}_{\bar{x}_A}/\bar{x}_A)$ = the proportional error for each replicate mean in Array A.

Raw hybridization values are used in Equation 3. When measurement error is the same proportion from assay to assay within each array, Equation 3 produces the same percentage error for all A/B ratios.

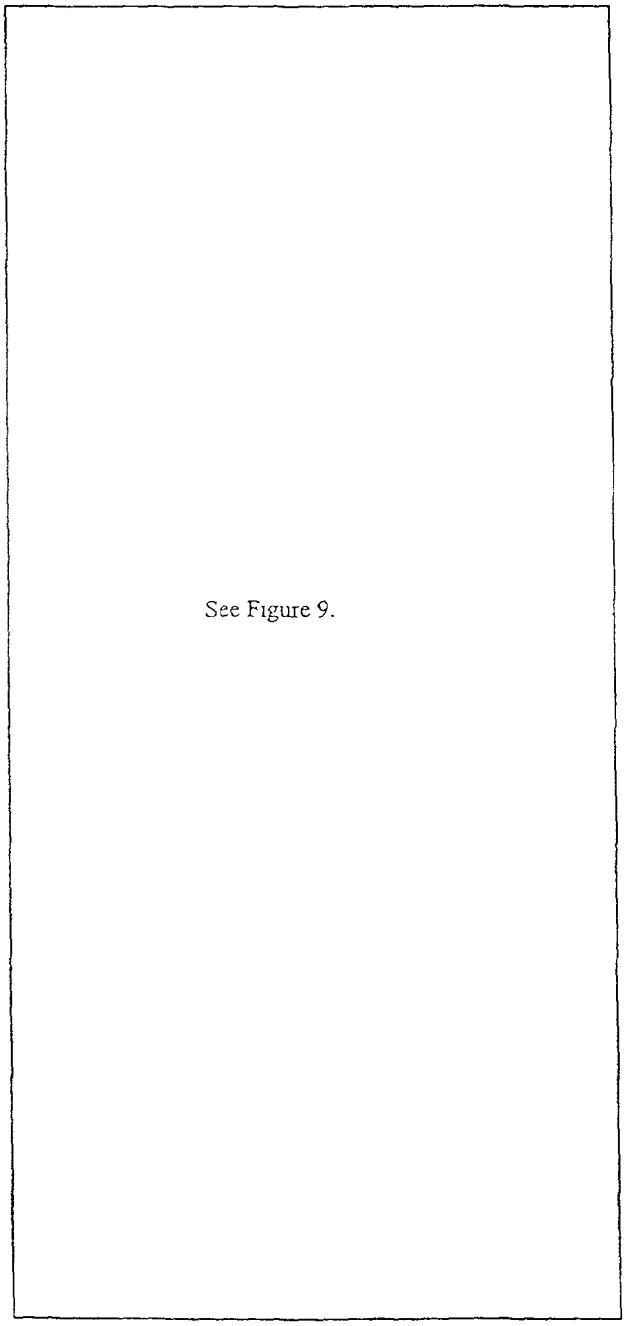
An analogous procedure is used when there are no replicates but an estimate of measurement error is available (e.g., from reference values or prior studies; see Appendix A)

The Process: Measurement Error Unkown

Option 2. **Generate confidence intervals for expression ratios using an estimate of measurement error derived from the distribution.** Confidence intervals can be developed for ratios, using an estimate derived from the variability of non-differentially expressed values.

We examine the variability of the middle 50% of log transformed hybridization ratios, which are assumed to be approximately distributed according to a Gaussian distribution. An estimate of the variability of ratios that are not differentially expressed is derived from this measure. This estimate is then used as discussed in Step 2 to assign confidence limits to all ratios (Fig. 7).

Figure 7: Confidence-based ratio evaluation. The assay at position 1,1 is compared to all other assays. Yellow indicates increased expression at 95% confidence ($p < .05$), and red at 99% confidence ($p < .01$).



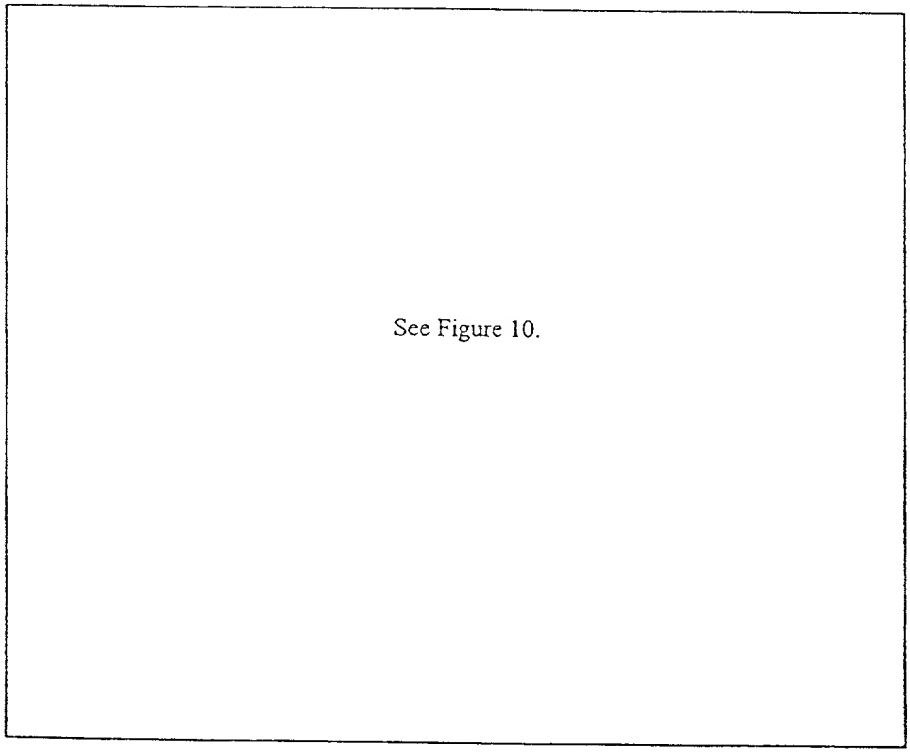
A Graphical Option

Model differential expression ratios using a Q-Q plot. The Q-Q plot is a member of the graphical statistics family. It maps frequency data to pdfs, in easily understood form. We use a Q-Q plot to model ratios of assays in one array divided by assays in another array (actually logs of the ratios of raw data). These ratios should present three partially overlapping distributions:

- values that do not differ across arrays (ratio 1:1);
- values that increase across arrays (ratio >1);
- values that decrease across arrays (ratio < 1);

The log values making up the distribution of values that do not differ should be normal. Therefore, we can use the central part of this distribution to model a complete distribution covering the range observed in the data. To the extent that observed values fail to lie within this distribution, they fall into the differentially expressed distributions (Figure 8).

Figure 8. Q-Q plot comparing the distribution of differential expression ratios (red line) to the Gaussian distribution (green line). Where observed values lie in close proximity to the straight line describing the expected value, they fall into the distribution of values that do not differ across arrays. Where the observed values deviate from the expected values, they fall into the distributions of differential expression.



The advantage of this procedure is that it can be performed with any ratios, even if we have no direct estimate of measurement error.

References

DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Yan, A.S. and Trent, J.M. Use of a cDNA microarray to analyse gene expression patterns in human cancer, *Nature Genetics* 14:457-460 (1996).

de Saizieu, A., Certa, U., Warrington, J., Gray, C., Keck, W. and Mous, J. Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays, *Nature Biotechnology* 16:45-48 (1998).

Nguyen, C., Rocha, D., Granjeaud, S., Baldit, M., Bernard, K., Naquet, P. and Jordan, B.R. Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones, *Genomics* 29:207-216 (1995).

Pietu, G., Alibert, O., Guichard, V., Lamy, B., Bois, F., Leroy, E., Mariage-Smason, R., Houlgatte, R., Soulare, P. and Auffray, C. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array, *Genome Research* 6:492-503 (1996).

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science* 270:467-470 (1995).

Shalon, D., Smith, S.J. and brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization, *Genome Research* 6:639-645 (1996)

United States Patent & Trademark Office

- Office of Initial Patent Examination

Application papers not suitable for publication

SN 09915813 Mail Date 07/25/01

☐ Non-English Specification

☑ Specification contains drawing(s) on page(s)_____ or table(s)✓

☐ Landscape orientation of text ☐ Specification ☐ Claims ☐ Abstract

☐ Handwritten ☐ Specification ☐ Claims ☐ Abstract

☐ More than one column ☐ Specification ☐ Claims ☐ Abstract

☐ Improper line spacing ☐ Specification ☐ Claims ☐ Abstract

☐ Claims not on separate page(s)

☐ Abstract not on separate page(s)

☐ Improper paper size -- Must be either A4 (21 cm x 29.7 cm) or 8-1/2"x 11"

- ☐ Specification page(s)_____ ☐ Abstract
- ☐ Drawing page(s)_____ ☐ Claim(s)

☐ Improper margins

- ☐ Specification page(s)_____ ☐ Abstract
- ☐ Drawing page(s)_____ ☐ Claim(s)

☐ Not reproducible

Reason

- ☐ Paper too thin
- ☐ Glossy pages
- ☐ Non-white background

Section

- ☐ Specification page(s)_____
- ☐ Drawing page(s)_____
- ☐ Abstract
- ☐ Claim(s)

☐ Drawing objection(s)

- ☐ Missing lead lines, drawing(s)_____
- ☐ Line quality is too light, drawing(s)_____
- ☐ More than 1 drawing and not numbered correctly
- ☐ Non-English text, drawing(s)_____
- ☐ Excessive text, drawing(s)_____
- ☐ Photographs capable of illustration, drawing(s)_____

United States Patent & Trademark Office

Office of Initial Patent Examination – Scanning Division

Application deficiencies found during scanning:

☐ Page(s)_____of_____ were not present for scanning. (Document title)

☐ Page(s)_____of_____were not present for scanning. (Document title)

The appendix is part of specs

☑ ***Scanned copy is best available.*** Drawings